

STIC-ILL

NPL
MIC QDI J77

From: Clow, Lori
Sent: Monday, June 09, 2003 2:35 PM
To: STIC-ILL
Subject: article request

I would like the following articles please:

Journal of Medicinal
Chemistry, American Chemical Society, vol. 38, No. 9, pp. 1431-1436,
(Apr. 28, 1995).

Brown, R.D. and Martin, Y.C., "Designing Combinatorial Library Mixtures
Using a Genetic Algorithm," Journal of Medicinal Chemistry, vol. 40, No.
15, American Chemical Society, 1997, pp. 2304-2313.

Gillet, V.J. et al., "The Effectiveness of Reactant Pools for Generating
Structurally-Diverse Combinatorial Libraries," J. Chem. Inf. Comput.

1997, v. 37, no. 4
p. 731-40

Kearsley, Simon K. et al., "Chemical Similarity Using Physiochemical
Property Descriptors," J. Chem. Inf. Comput. Sci., vol. 36, No. 1,
American Chemical Society, 1996, pp. 118-127.

Leland, B.A. et al., "Managing the Combinatorial Explosion," J. Chem.
Inf. Comput. Sci., vol. 37, No. 1, American Chemical Society, 1997, pp.
62-70.

Lewis, R.A. et al., "Similarity Measures for Rational Set Selection and
Analysis of Combinatorial Libraries: The Diverse Property-Derived (DPD)
Approach," J. Chem. Inf. Comput. Sci., vol. 37, No. 3, American Chemical
Society, 1997, pp. 599-614.

Martin, E.J. and Critchlow, R.E., "Beyond Mere Diversity: Tailoring
Combinatorial Libraries for Drug Discovery," Journal of Combinatorial
Chemistry, vol. 1, No. 1, American Chemical Society, Jan. 1999, pp.
32-45.

Agrafiotis, D.K. and Lobanov, V.S., "Ultrafast Algorithm for Designing
Focused Combinatorial Arrays," J. Chem. Inf. Comput. Sci., vol. 40, No.
4, American Chemical Society, Jun. 16, 2000, pp. 1030-1038.

Good, A.C. and Lewis, R.A., "New Methodology for Profiling Combinatorial
Libraries and Screening Sets: Cleaning Up the Design Process with
HARPick," J. Med. Chem., vol. 40, No. 2, American Chemical Society,
1997, pp. 3926-3936.

Jamois, E.A. et al., "Evaluation of Reagent-Based and Product-Based
Strategies in the Design of Combinatorial Library Subsets," J. Chem.
Inf. Comput. Sci., vol. 40, No. 1, American Chemical Society, 2000
(pub'd on Web Dec. 9, 1999), pp. 63-70.

Thanks

Lori A. Clow, Ph.D.
Patent Examiner, Art Unit 1631
Crystal Mall 1, Room 12-D-08
Mailbox 12-D-01
(703)-306-5439

3/5/3 (Item 1 from file: 34)

06008932 **Genuine Article#:** XN700 **Number of References:** 26

The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries

Author: Gillet VJ (REPRINT) ; Willett P; Bradshaw J

Corporate Source: UNIV SHEFFIELD,KREBS INST BIOMOLEC RES/SHEFFIELD S10 2TN/S YORKSHIRE/ENGLAND/ (REPRINT); UNIV SHEFFIELD,DEPT INFORMAT STUDIES/SHEFFIELD S10 2TN/S YORKSHIRE/ENGLAND/; GLAXO WELLCOME RES & DEV LTD,/STEVENAGE SG1 2NY/HERTS/ENGLAND/

Journal: JOURNAL OF CHEMICAL INFORMATION AND COMPUTER SCIENCES , 1997 , V 37 , N4 (JUL-AUG) , P 731-740

ISSN: 0095-2338 **Publication date:** 19970700

Publisher: AMER CHEMICAL SOC , 1155 16TH ST, NW, WASHINGTON, DC 20036

Language: English **Document Type:** ARTICLE

Abstract: Current approaches to the design of combinatorial libraries assume that structural diversity in the reactant pools corresponds to structural diversity in the combinatorial libraries that result from reacting these pools together. In experiments with three different published libraries, dissimilarity-based compound selection (DBCS) is applied at two levels. First, the DBCS algorithm is applied at the reactant level, a library is built, and its diversity is measured. Second, the DBCS algorithm is applied to the full set of products generated by enumeration of all the reactants and the diversity of the subset is measured. Results show that reactant-based selection, which attempts to maximize diversity in the pools, results in noticeably less diverse libraries than if the selection is performed at the product level. Experiments are reported to estimate the upperbound to diversity achievable using DBCS, and it appears that DBCS is very effective at finding maximally diverse subsets. However, applying DBCS selection at the product level is synthetically inefficient since it does not result in a combinatorial Library. We thus describe a genetic algorithm for selecting combinatorial libraries from the fully enumerated products and demonstrate that these libraries are significantly more diverse than those generated using reactant-based selection.

SciSearch(R) Cited Ref Sci (Dialog® File 34): (c) 2003 Inst for Sci Info. All rights reserved.

© 2003 The Dialog Corporation

The Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries

Valerie J. Gillet,^{*,†} Peter Willett,[†] and John Bradshaw[‡]

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom, and GlaxoWellcome Research and Development Limited, Gunnels Wood Road, Stevenage, SG1 2NY, United Kingdom

Received January 23, 1997[®]

Current approaches to the design of combinatorial libraries assume that structural diversity in the reactant pools corresponds to structural diversity in the combinatorial libraries that result from reacting these pools together. In experiments with three different published libraries, dissimilarity-based compound selection (DBCS) is applied at two levels. First, the DBCS algorithm is applied at the reactant level, a library is built, and its diversity is measured. Second, the DBCS algorithm is applied to the full set of products generated by enumeration of all the reactants and the diversity of the subset is measured. Results show that reactant-based selection, which attempts to maximize diversity in the pools, results in noticeably less diverse libraries than if the selection is performed at the product level. Experiments are reported to estimate the upperbound to diversity achievable using DBCS, and it appears that DBCS is very effective at finding maximally diverse subsets. However, applying DBCS selection at the product level is synthetically inefficient since it does not result in a combinatorial library. We thus describe a genetic algorithm for selecting combinatorial libraries from the fully enumerated products and demonstrate that these libraries are significantly more diverse than those generated using reactant-based selection.

INTRODUCTION

The last few years have seen an explosive growth in the use of combinatorial methods for the creation of extremely large libraries of structurally-diverse molecules, from which it has proved possible to identify biologically-active molecules far more rapidly than is possible using conventional approaches to drug discovery.¹⁻⁴ The effectiveness of the approach is crucially dependent on the building blocks, or *reactants*, that are used as the input to the combinatorial synthesis of the final *products* since there are generally far more reactants available than can actually be used in practice.⁵ For example, peptoids are polymers of N-substituted glycine that have a peptide backbone but with side chains attached at the amide nitrogen instead of the α -carbon.⁶ Peptoids are synthesized by incorporating the side chains from amines. Public databases of chemical structures, such as the *Available Chemicals Directory*, may contain many hundreds of different, readily-available amines, thus permitting the synthesis of libraries containing extremely large numbers of different molecules, even if attention is restricted to tri- and tetrapeptoids.

Techniques for combinatorial synthesis have developed rapidly, and it is now possible to synthesize extremely large numbers of compounds in single combinatorial experiments. Combinatorial synthesis, therefore, is an efficient way of providing compounds for high throughput screening for the discovery of new leads. However, it is the rate at which compounds can be screened that is the limiting step in combinatorial chemistry. One way of increasing throughput is to synthesize and screen compounds as mixtures rather

than as discrete compounds. Although this method does allow a large number of compounds to be screened, there are several problems associated with the handling of mixtures. These include the difficulty of assessing the quality of the mixtures to determine whether all intended compounds have actually been synthesized and the possibility that a positive screening result may in fact be the result of the synergy of several structures so that no activity is found when the compounds are deconvoluted and screened independently. The problems associated with synthesizing and screening mixtures are avoided by the parallel synthesis and testing of discrete compounds. Since a much smaller number of compounds can be screened, it is then necessary to use compound selection in order to reduce the number of compounds available for testing. There has, therefore, been much interest in techniques for the selection of sets of dissimilar reactants from existing chemical databases, so that the compounds that are generated cover a wide range of structural types.⁵⁻¹⁴ It is assumed that if it is possible to identify maximally-diverse (or, more realistically, near maximally-diverse) sets of reactants, then their use will result in the generation of a maximally-diverse combinatorial library of products. If this assumption is correct, it will permit the full exploration of the potential structural space even though only a relatively small number of compounds are actually synthesized and tested. In what follows, the assumption that diversity at the reactant level reflects diversity at the product level is referred to as the *diversity hypothesis*.

The assumption that a diverse set of products will result from a diverse set of reactants has not yet been tested experimentally owing to the sheer numbers of compounds that are involved. For example, the tripeptoid library investigated by Martin *et al.* was based on no less than 721 primary amines (and 1133 carboxylic acid and acid-chloride

* Author to whom correspondence should be sent. E-mail: v.gillet@sheffield.ac.uk.

[†] University of Sheffield.

[‡] GlaxoWellcome Research and Development Limited.

[®] Abstract published in *Advance ACS Abstracts*, June 15, 1997.

amino-terminal capping groups) but used only 18 members of the reactant pool.⁶ This paper reports a quantitative examination of the validity of the diversity hypothesis. We show that the diversity hypothesis is incorrect and that selection of diverse reactants does not result in maximum diversity in product space. We then report on a more effective method that we have developed for selecting reactants by analyzing product space. While still not optimal in terms of maximizing a quantitative index of structural diversity, this approach provides a noticeably better solution than existing methods for the selection of reactants in order to maximize the diversity of combinatorial libraries.

TESTING THE DIVERSITY HYPOTHESIS

Theoretical Background. Consider a combinatorial library, c , that is synthesized from reactants contained in two reactant pools, r_1 and r_2 , of sizes n_1 and n_2 , respectively (in the following, we consider only dimer libraries for the purpose of simplicity). These two reactant pools have previously been selected as representing diverse subsets of two larger potential-reactant pools, R_1 and R_2 , of sizes N_1 and N_2 , respectively, using some quantitative subset-selection procedure. Let C be the combinatorial library that would have been generated from all possible combinations of R_1 and R_2 if the subset-selection procedure had not been used. Thus, c and C contain n_1n_2 and N_1N_2 dimers, respectively. The same subset-selection procedure that was used to create the reactant pools r_1 and r_2 , (i.e., that was used to identify the n_1 most dissimilar molecules in R_1 and the n_2 most dissimilar molecules in R_2) is used to identify the most dissimilar n_1n_2 molecules from amongst the N_1N_2 molecules in C . This subset is referred to subsequently as library L_d^* . The construction of the libraries c and L_d^* is illustrated in Figure 1.

Let $D(X)$ be a function that returns a value describing the diversity of a set of molecules, X . Then the diversity hypothesis would suggest that c is comparable in structural diversity to L_d^* , i.e., that

$$D(L_d^*) = D(c)$$

This is actually a limiting case since it is easy to prove that

$$D(L_d^*) \geq D(c)$$

by contradiction. Assume that the converse is true and that there is thus a subset of size n_1n_2 from c that has a greater level of diversity (however this is defined) than the subset of the same size from C . Now both c and L_d^* are subsets of C , i.e., $c \subset C$ and $L_d^* \subset C$, and thus every member of c must also be in C . However, if the subset L_d^* is defined to be the maximally-diverse subset that can be generated from C then, of necessity,

$$D(L_d^*) \geq D(c)$$

This contradicts the original assumption, which must thus be false. These arguments demonstrate that it is possible for a subset-selection procedure to identify a subset that is equal in diversity to that of a fully enumerated library but that it is never possible (in accordance with intuition) to identify one that is superior. In fact, given that L_d^* is selected from the fully enumerated library it is very likely that $D(L_d^*) > D(c)$.

Thus far, it has been assumed that the subsets r_1 , r_2 , and L_d^* are maximally diverse, however, this is unlikely to be achieved in practice. Selection of the maximally-diverse subset is computationally infeasible since it requires evaluation of

$$\frac{N!}{n!(N-n)!}$$

subsets, where a subset of n compounds is selected from a library containing N compounds. Let $D(\max)$ be the maximum diversity that is possible for a subset of C of size n_1n_2 . It is known that $D(c)$ and $D(L_d^*)$ are unlikely to be equal to $D(\max)$. It is also possible to find the most similar subset of compounds of a collection. Let this subset be L_s^* with diversity $D(L_s^*)$. Any algorithm for the selection of the most similar subset is also likely to be suboptimal, and if $D(\min)$ is the maximally-similar subset, then it is likely that $D(L_s^*) > D(\min)$. Let $D(L_r^*)$ be the diversity of a subset that is selected at random from C . The subsets selected from C are summarized in Figure 2 and are referred to as libraries of compounds; however, they are most unlikely to represent combinatorial libraries. Assuming that the library c is of greater diversity than a library selected at random, the relative ordering of the diversities of the libraries is then expected to be

$$D(\min) < D(L_s^*) < D(L_r^*) < D(c) < D(L_d^*) < D(\max)$$

In the first set of experiments described in this paper, $D(L_d^*)$, $D(L_r^*)$, $D(L_s^*)$, and $D(c)$ were measured for three different libraries in an attempt to test the diversity hypothesis. Given subset-selection procedures that are not guaranteed to be optimal, the question at issue when considering the diversity hypothesis is whether it is possible to achieve (near)-equally diverse subsets, and whether subset-selection at both the reactant level and at the product level is significantly more effective than selecting a subset at random. If $D(\max)$ and $D(\min)$, the upper- and lowerbounds on diversity, respectively, are known, then a more quantitative understanding of the difference between $D(L_d^*)$, $D(c)$, and $D(L_r^*)$ can be determined. However, $D(\max)$ and $D(\min)$ cannot be measured directly, and the second set of experiments was designed to provide estimates for these limiting values. The final experimental section describes the new algorithm we have developed for selecting combinatorial libraries from product space so as to maximize diversity.

Experimental Details. Two procedures are required to demonstrate the validity of the diversity hypothesis: a method of selecting maximally diverse subsets that can be applied to the reactant pools and to the product library C and a method of quantifying the inherently qualitative concept of "molecular diversity".

Dissimilarity-based compound selection (DBCS)¹⁰⁻¹⁴ involves the identification of the maximally-diverse subset of size n from a database of size N (where, typically, $n \ll N$). The DBCS algorithm involves summing the dissimilarities of every molecule with all of the other molecules in the set. The first molecule to be selected is that which is most dissimilar from all of the others, i.e., it has the greatest sum of dissimilarities of all the molecules. The second molecule to be selected is that which is most dissimilar to the first. The third molecule is that which is most dissimilar to the

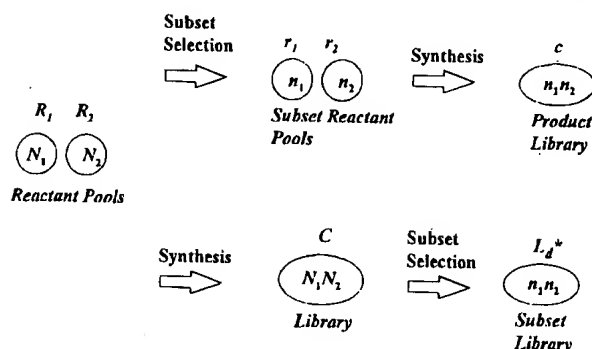


Figure 1. Subset selection can be performed at either the reactant level or at the product level. The diversity hypothesis assumes that diversity at the reactant level reflects diversity at the product level.

first and the second and so on. The DBCS method used here for the selection of diverse reactants (the pools r_1 and r_2) and diverse products (the library L_d^*) is that described by Holliday *et al.*¹⁴ and is based on the cosine coefficient. This implementation of the DBCS algorithm can be applied to sets of molecules using any structural representation that is described by a vector. We have chosen to use Daylight fingerprints,¹⁵ containing 1024 bits, for most of the experiments reported here although we have also briefly tested the hypothesis for another representation that is described later.

The DBCS algorithm can be modified easily to select the most similar subset of molecules from a collection. Thus, a *similarity-based compound selection* (SBCS) algorithm was implemented by measuring pairwise similarities, rather than dissimilarities, and selecting molecules that are most similar, rather than most dissimilar, to those already selected. The SBCS algorithm was applied to the fully enumerated library, C , to generate the subset of most similar compounds, referred to as library L_s^* .

In calculating the diversities of the various sets of compounds, we have followed previous workers^{5,6,9} in assuming that the diversity of a set of molecules can be determined from the intermolecular structural dissimilarities for that dataset. Specifically, we have used the diversity measure described by Turner *et al.*,¹⁶ which is the mean intermolecular dissimilarity when averaged over all the pairs of molecules in a dataset and which provides an easily calculable single-valued representation of the diversity of molecules in a dataset.

The DBCS algorithm was first used to identify the most dissimilar $n_1 n_2$ molecules from amongst the $N_1 N_2$ molecules in C ; the selected molecules form the subset library L_d^* . The algorithm was then used to create the reactant pools r_1 and r_2 by identifying the n_1 most dissimilar molecules from amongst the N_1 molecules in R_1 and the n_2 most dissimilar molecules from amongst the N_2 molecules in R_2 ; the combination of these two reactant pools forms the combinatorial library c . The diversities of L_d^* and c were then calculated and compared, as detailed in Figure 3. Random subsets, L_r^* , and subsets of similar compounds, L_s^* , were also selected for comparison by analogous procedures.

The selection of a set of n molecules from a database of N molecules using our DBCS algorithm has a time complexity of order $O(nN)$.¹⁴ The creation of the library L_d^* requires the selection of $n_1 n_2$ diverse molecules from the $N_1 N_2$ molecules in C (step 2 in Figure 3) and hence has a time

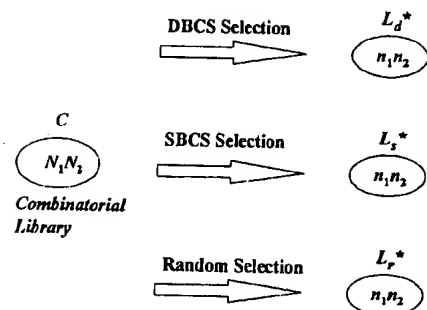


Figure 2. Different libraries selected from the fully enumerated combinatorial library C : L_d^* is the library selected by applying DBCS; L_s^* is the library selected by applying SBCS; and L_r^* is a library selected at random.

complexity of $O(n_1 n_2 N_1 N_2)$. Similarly, the creation of the combinatorial library c requires the generation of the two diverse reactant pools, r_1 and r_2 (steps 4 and 5 in Figure 3) and hence has a total time complexity of order $O(n_1 N_1) + O(n_2 N_2)$. The remaining steps in Figure 3 have complexities of $O(N_1 N_2)$ for step 1, $O(n_1 n_2)$ for steps 3, 6, and 7, and $O(1)$ for step 8. Step 2 hence dominates the computation, and the procedure thus has an overall complexity of $O(n_1 n_2 N_1 N_2)$. In most of the experiments reported below, both n_1 and n_2 were 40 and both N_1 and N_2 were 400, with the selection of 1600 compounds from 160 000 taking approximately 2.6 h on a Silicon Graphics R10000 workstation.

The experiments involved three different combinatorial library systems: an amide library that was built by coupling carboxylic acids to primary amines by forming a peptide bond; two related libraries based on benzoic acid as a template; and a library based on Kemp's acid. It should be noted here that the experiments are "paper" experiments designed to test the diversity hypothesis and they are not real synthetic experiments. The libraries varied from having no common substructural core, in the first example, through having a small core substructure in the benzoic acid examples, to having a large core substructure where the substituents are generally relatively small compared to the core itself, in the Kemp's acid example. In each case, the Daylight Toolkit¹⁶ was used to develop software to perform the required "reaction" between reactants in different pools in order to enumerate the libraries. In all of the experiments, C contained 160 000 products and the sizes of L_d^* , L_s^* , L_r^* , and c were varied.

RESULTS

Amide Library. The amide combinatorial library, C , was enumerated by coupling carboxylic acids to primary amines through the formation of a peptide bond, see Figure 4. A pool of amine reactants was formed by choosing a random set of 400 molecules from the primary amines that are present in the World Drug Index (WDI).¹⁷ Similarly, a pool of carboxylic acid reactants was formed by choosing from the same file a random set of 400 molecules that contain a single carboxylic acid group. Software was developed to join combinations of amines and carboxylic acids by forming peptide bonds.

The full library, C , of size 160 000, was constructed by joining all 400 amines to all 400 carboxylic acids. In the first experiments, a diverse subset, L_d^* , containing 1600 molecules was selected from C using the DBCS algorithm¹⁴

1. Create the N_1N_2 products in library C by combining each of the N_1 reactants in R_1 with each of the N_2 reactants in R_2 .
2. Create the library L_d^* by selecting the n_1n_2 most diverse products from C .
3. Calculate the sums of dissimilarities for all pairs of products in L_d^* , and hence the diversity $D(L_d^*)$.
4. Create r_1 by selecting the n_1 most diverse reactants from R_1 .
5. Create r_2 by selecting the n_2 most diverse reactants from R_2 .
6. Create the n_1n_2 products in library c by combining each of the n_1 reactants in r_1 with each of the n_2 reactants in r_2 .
7. Calculate the sums of dissimilarities for all pairs of products in c , and hence the diversity $D(c)$.
8. Compare $D(L_d^*)$ with $D(c)$.

Figure 3. Procedure for generating libraries c and L_d^* that can then be used to test the diversity hypothesis.

Table 1. Test of the Diversity Hypothesis Using an Amide Library^a

$D(L_d^*)$	$D(c)$	$D(L_r)$	$D(L_s^*)$
0.652	0.596	0.508 (0.003)	0.132
0.651	0.589	0.509 (0.004)	0.134
0.651	0.594	0.510 (0.004)	0.132

^a $D(L_r^*)$ gives the mean diversity and standard deviation (in brackets) for 1000 subsets chosen at random.

Table 2. Effect of Library Size, $\#(c)$, on Diversity Using an Amide Library

$\#(c)$	$D(L_d^*)$	$D(c)$	$D(L_r^*)$	$D(L_s^*)$
1600	0.652	0.596	0.509 (0.003)	0.132
900	0.658	0.594	0.509 (0.006)	0.125
400	0.663	0.596	0.509 (0.015)	0.127
100	0.674	0.582	0.513 (0.059)	0.100

and Daylight fingerprint representations of the molecules. The same procedure and structural representation was then used to select 40 diverse amines and 40 diverse acids from the two pools of 400 reactants; the selected acids and amines were used to generate the combinatorial library, c , of 1600 amides. The diversity hypothesis was then tested by comparing the $D(c)$ and $D(L_d^*)$ values using the procedure shown in Figure 3. $D(L_s^*)$ and $D(L_r^*)$ were also measured. The entire procedure was repeated three times, using different randomly-chosen pools of amine and carboxylic acid reactants.

The results are shown in Table 1, where it can be seen that the relative ordering of the libraries is as expected: $D(L_s^*) < D(L_r^*) < D(c) < D(L_d^*)$. A more diverse library of compounds is thus generated by performing compound selection at the product level rather than at the reactant level, and selection at both the product and reactant levels results in more diverse libraries than selection of a subset of compounds at random. The distribution of diversities is not symmetrical; that is, the diversity of a library selected at random is closer to the diversity of the library selected using the DBCS algorithm than to the diversity of the library selected using the SBCS algorithm. This is because of the nature of a combinatorial library, where a small number of reactants is used to generate a large number of products. There are many subsets of very similar compounds, for example, a subset where the carboxylic acid component is constant and variation is seen in the amine component only.

Further runs were carried out for different sizes of subsets; specifically libraries of 1600, 900, 400, and 100 compounds were selected that correspond to reactant subset sizes of 40, 30, 20, and 10, respectively. The results of these runs are shown for an amide library in Table 2, where it will be seen that the $D(L_d^*)$ values increase as the size of the library decreases. This is because an algorithm for DBCS will initially tend to select molecules from "around the edges" and then move toward the "center" of the dataset once its periphery has been fully explored. The first molecules selected, which will be those in a very small subset, will hence tend to be more dissimilar to each other than those selected later, and since diversity is measured as the mean intermolecular dissimilarity averaged over all molecules in the set, the diversity increases as the library size decreases, with the resulting trend that is shown in Table 2. The mean diversities of the randomly chosen subsets do not vary much but the range of values increases as the subsets decrease in size. Less variation is seen in the $D(c)$ values than the $D(L_d^*)$ values, with the 1600, 900, and 400 subsets having very similar diversities. However, the smallest combinatorial library of 100 molecules shows an unexpected decrease in diversity that is not statistically significantly different from the diversity of subsets chosen at random. In selecting the most diverse reactants no account is taken of the diversity of one reactant subset relative to the other, and it could be that, although diversity within a subset is maximized, the subset as a whole contains molecules that are similar to those found in the other subset. This

Table 3. Effect on $D(c)$ When Reactants Are Chosen To Maximize the Diversity of Both Reactant Pools Taken Together

#(c)	$D(c)$	#(c)	$D(c)$
1600	0.599	400	0.607
900	0.602	100	0.607

Table 4. Test of the Diversity Hypothesis Using an Amide Library^a

$D(L_d^*)$	$D(c)$	$D(L_r^*)$
0.449	0.346	0.133 (0.003)

^a Dissimilarity comparisons are made using structural features. L_d^* contains the 1600 compounds that result from performing DBCS on the product library C , and c contains 1600 compounds that result from using DBCS to select 40 diverse reactants from each pool and then enumerating the products. $D(L_r^*)$ gives the mean diversity and standard deviation (in brackets) for 1000 subsets of size 1600 chosen at random.

assumption was tested by altering the way in which reactants are selected. Table 3 shows the results obtained when reactants are selected alternately and the diversity of the two subsets together is maximized. In this case, the $D(c)$ values also increase as the size of the subsets decrease.

Table 4 shows the results obtained when a different kind of representation was used to measure the dissimilarity between molecules. In this case, the molecules were represented by a number of structural features that might better relate to biological activity. Each structure is represented by the following six different features: counts of hydrogen bond donors, hydrogen bond acceptors, rotatable bonds, and aromatic rings and the physical property values of molecular weight and the $^2\kappa_a$ shape descriptor. A hydrogen bond donor is defined as any heteroatom that carries at least one hydrogen. A hydrogen bond acceptor is defined as a heteroatom, excluding the halogens, aromatic oxygen, sulfur, and pyrrole nitrogen and the higher oxidation levels of nitrogen, phosphorus, and sulfur. (All of the compounds used in the experiments have neutral charge.) The feature values were standardized to fit into the range of 0.1. This is achieved by finding the maximum and minimum values over the whole library of molecules. The normalized value for each feature, x , in each molecule is then given by

$$\frac{x - \min}{\max - \min}$$

$D(c)$ and $D(L_d^*)$ were calculated using the DBCS algorithm as before. These results again show that a more diverse library results if selection is performed at the product level rather than at the reactant level, demonstrating that our results are not dependent on the particular structural representation that is used.

3-Amino-5-hydroxybenzoic Acid Library. Dankwardt *et al.*¹⁸ have discussed the generation of combinatorial libraries in which the core molecule is 3-amino-5-hydroxybenzoic acid. In one of their experimental schemes, the carboxylic acid serves as a handle onto which the solid support is attached and diversity is incorporated onto the two remaining functional groups. Two different reactions were simulated in the present study, both involving the substitution of carboxylic acids onto the amine group.

In the first set of experiments the hydroxyl group was substituted by sulfonyl chlorides, as shown in Figure 5a. In

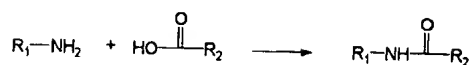
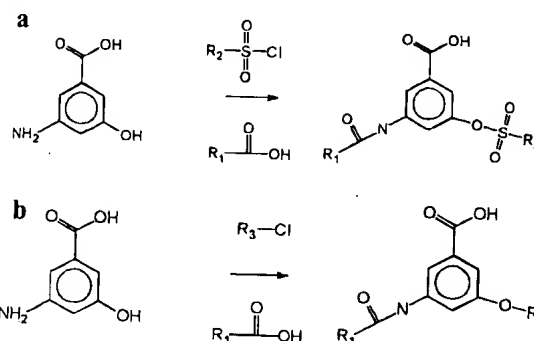

Figure 4. The amide library.

Figure 5. (a) The benzoic acid library with carboxylic acids and sulfonyl chlorides as substituents. (b) The benzoic acid library with carboxylic acids and chlorides as substituents.

Table 5. Test of the Diversity Hypothesis Using a 3-Amino-5-hydroxybenzoic Acid Library

#(c)	$D(L_d^*)$	$D(c)$	$D(L_r^*)$	$D(L_s^*)$
(a) Use of Sulfonyl Chlorides				
1600	0.408	0.388	0.358 (0.002)	0.130
900	0.410	0.387	0.359 (0.002)	0.117
400	0.413	0.384	0.360 (0.004)	0.105
100	0.421	0.373	0.365 (0.020)	0.093
(b) Use of Chlorides				
1600	0.435	0.411	0.390 (0.001)	0.167
900	0.437	0.410	0.391 (0.001)	0.143
400	0.440	0.409	0.391 (0.003)	0.114
100	0.448	0.404	0.396 (0.009)	0.086

the second set of experiments, the hydroxyl group was substituted by chloride-containing substituents (alkyl halides in the original paper but modified here to be any compound containing a single chlorine atom), as shown in Figure 5b. Reactant pools were created by selecting 400 molecules at random from the SPRESI database.¹⁹ One pool consisted of molecules containing a single carboxylic acid group; another consisted of molecules containing a sulfonyl chloride group; and a third pool consisted of molecules containing a single chlorine atom. Libraries containing 160 000 molecules were enumerated from the reactant pools, and 1600-member subsets c , L_d^* , L_r^* , and L_s^* were then generated as described previously.

The $D(L_d^*)$, $D(c)$, $D(L_r^*)$, and $D(L_s^*)$ values for these datasets are listed in Table 5 (parts a and b), where the same behavior is observed as was the case with the amide libraries; that is, reactant-based selection lies between random selection and product-based selection. Note that all of the diversity values here are lower than for the amides, presumably due to the common core substructure that is present in all these products, and similar comments apply to the Kemp's library discussed below.

Kemp's Acid Library. Kocis *et al.*²⁰ have demonstrated the use of Kemp's acid as a useful building block for the preparation of synthetic receptors. The core molecule used is shown in Figure 6. Two pools of reactants were used at substitution positions R_1 and R_2 , with each of these reactant pools containing 400 carboxylic acids chosen at random from carboxylic acids extracted from the SPRESI database. The libraries C , c , L_d^* , L_r^* , and L_s^* were generated as described previously, with the results shown in Table 6 being compa-

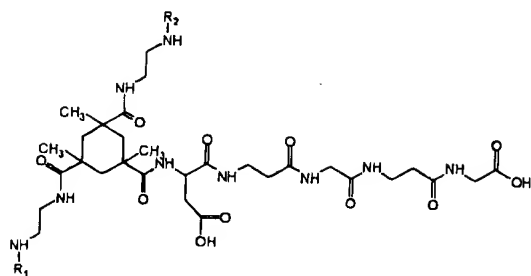


Figure 6. The Kemp's acid library.

Table 6. Test of the Diversity Hypothesis Using a Library Derived from Kemp's Acid

#(c)	$D(L_d^*)$	$D(c)$	$D(L_r^*)$	$D(L_r^*)$
1600	0.435	0.402	0.380 (0.001)	0.169
900	0.437	0.401	0.380 (0.001)	0.145
400	0.440	0.403	0.381 (0.003)	0.116
100	0.448	0.396	0.386 (0.012)	0.106

able to those obtained with the amide and 3-amino-5-hydroxybenzoic acid libraries.

ESTIMATING THE UPPERBOUND OF DIVERSITY

The results obtained above suggest that selection at the reactant level is less effective than selection at the product level, this implying that the diversity hypothesis is not correct. However, it is difficult to quantify the differences in diversities without knowing the bounds on diversity for subsets selected from a library. This section describes experiments that attempted to determine $D(\max)$ in order to evaluate the effectiveness of reactant-based selection compared to product-based selection.

Three different methods were used in order to estimate how near-optimal the subset selection method is, *i.e.*, how close $D(L_d^*)$ is to $D(\max)$. These were as follows: measuring the variation in the diversities of subsets of a given size chosen at random; comparing the DBCS result with a genetic algorithm that was designed to maximize the diversity of subsets; and using extreme value methods to estimate the end point of the distribution of diversities. The experiments were applied to the problem of selecting 40 diverse molecules from a set of 400 carboxylic acids extracted at random from the WDI database. The diversities that result in each experiment were compared with the diversity produced by the DBCS method, which gave a value of 0.698, *i.e.*, $D(L_d^*) = 0.698$.

Standard Deviation. The performance of the DBCS algorithm was estimated by generating many subsets at random in an empirical test of goodness of the heuristic.²¹ Three million subsets of size 40 were selected at random and their diversities measured using the Daylight fingerprint representation of molecular structure. The mean diversity was calculated as 0.618 with a standard deviation of 0.016. The subset selected by the DBCS method has diversity of 0.698 which is 5.0 standard deviations above the mean. If it is assumed that the subset diversities are normally distributed, then the DBCS-selected subset is superior to no less than 99.999 999 999 9% of all possible subsets. While this result says nothing about the difference between $D(\max)$ and $D(L_d^*)$, it does demonstrate that the DBCS algorithm is able to generate subsets with very high diversities using our chosen diversity index.

Genetic Algorithm. A genetic algorithm (GA)²² was used to explore the diversity space of different subsets generated from the library. A GA is the computational analogue of Darwinian evolution. Potential solutions to a problem are encoded in a population of chromosomes which are linear representations of the problem that is to be solved and which are scored using a fitness function. New populations are developed by performing genetic-like operations of mutation and crossover on some members of the existing population. Higher scoring individuals have a higher probability of passing their genes into the new populations. The new chromosomes are scored and the GA iterates, usually until it has converged on a solution. GAs have previously been applied to many problems in computational chemistry,²³ including the design of combinatorial libraries targeted at one particular biological assay.⁵ Here, the GA was developed to generate diverse subsets of molecules.

Each chromosome in the GA represented a particular 40-member subset with each element of a chromosome representing a molecule in the library. Thus, a chromosome contained 40 integers in the range 1–400 corresponding to the 400 carboxylic acids that are available for selection. The GA was initialized with a population of 50 chromosomes that represented 50 different randomly chosen subsets. The genetic operators crossover and mutation were applied to evolve new subsets, ensuring that 40 unique molecules were present in each of the subsets. Mutation involved changing an element to a new randomly chosen element that represented a molecule that was not already contained in the subset. One-point crossover was modified so that an element was only exchanged if the molecule represented by the new element did not already exist in the chromosome. The chromosomes were scored by calculating the diversity of the subsets of molecules they represented, so that the fittest chromosomes were those that represented the most diverse subsets. The fitness function therefore attempts to maximize the diversity of the subsets represented by the chromosomes.

Very many experiments were carried out to identify the best parameter values for the GA. A high rate of mutation was performed relative to crossover (3:1) in order to prevent the GA from converging on a local, rather than a global, maximum. The GA iterated until it had converged on a solution, *i.e.*, until it could not find a better subset. Once this condition had been reached, the GA was rerun with a new starting population consisting of the best 10% of the chromosomes from the previous run, with the remaining chromosomes initialized to random subsets. After 100 runs of the GA, the best result obtained was a subset with diversity exactly equal to that found by the DBCS method (0.698). Thus, the GA was unable find a more diverse subset than is found using the DBCS subset-selection algorithm, despite searching through the diversity space of a large number of subsets selected from the library. This suggests that the greedy algorithm approach embodied in DBCS provides an extremely effective heuristic for selecting diverse subsets.

Extreme Value Methods. Extreme value methods have been used as ways of estimating the extreme behavior of a process on the basis of an observed independent distribution.²⁴ They have been applied in engineering situations where structures such as oil-rigs typically fail, *e.g.*, overturn, because of the occurrence of extreme values of a single environmental process or a critical extreme combination of constituent variables, such as sea surface waves and winds.

The essential problem is one of extrapolation: to estimate some unknown distribution function beyond the end of the known observations or, in other words, to estimate the end point of a distribution in a finite population. Extreme value methods can be applied to the problem of estimating the maximum diversity of a subset of a given size extracted from a combinatorial library. A finite distribution of diversities exists for all possible subsets of molecules of a given size selected from a library. The known observations that have already been measured for such a distribution include $D(L_d^*)$, $D(L_r^*)$, and $D(L_s^*)$, where L_d^* is the subset with the largest calculated diversity. It is possible to generate any number of observations, for example, by selecting subsets at random and measuring their diversities as done previously and then using the observed distribution to predict the end point of the distribution of all possible subsets.

The extreme value method applied here was the Generalised Pareto Distribution (GPD)²⁵ which has been widely used to model the upper tails of distributions. The GPD method was applied to subset selection in order to estimate the diversity of the maximally-diverse subset. In general, observations with the largest values are expected to provide the most valuable information on the end point in the GPD method, and, hence, a number of independent observations that are close to the best observation, $D(L_d^*)$, are required. Data points close to $D(L_d^*)$ were generated by seeding a subset with five randomly chosen structures. The remaining 35 structures in the 40-member subset were chosen by applying the DBCS algorithm to the full set of 400 carboxylic acids. This process was repeated 100 times to give 100 data points. None of these subsets was more diverse than the DBCS selected subset. The DBCS observation was also included in the sample. The GPD method involves sampling values in the distribution above a threshold. The best estimate for the maximum diversity using this method was 0.699 with a standard deviation of 0.0002, which is almost identical to the best result found using the GA. Even allowing for five standard deviations above the mean, this still gives a value for $D(\max)$ of only 0.700, thus again suggesting the $D(L_d^*)$ is a fair approximation to $D(\max)$.

SELECTING COMBINATORIAL LIBRARIES FROM PRODUCTS

The experiments in the previous section suggest that the DBCS algorithm is very effective at finding a subset that is very close in diversity to the maximally diverse subset, i.e., $D(L_d^*) \approx D(\max)$. For all of the libraries tested above, the mean diversity, $D(L_r^*)$, and standard deviation for 1000 randomly chosen subsets of C of size 1600 represent an approximate lowerbound to the effectiveness of reactant-based selection procedures for these classes of structures. It is clear that the $D(L_r^*)$ values are a large percentage of both the $D(c)$ and $D(L_d^*)$ values. It is certainly the case that the $D(c)$ values are closer to the $D(L_d^*)$ values than they are to the $D(L_r^*)$ values; however, it is also clear that there is still considerable scope for improving the diversity of the products and that the correctness of the diversity hypothesis is not supported. This suggests that it would be beneficial to develop selection methods that operate in product space. However, although greater diversity can be achieved by applying DBCS at the product level, this technique is synthetically inefficient since the subsets of molecules do

a	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9
x_1	x_1y_1	x_1y_2	x_1y_3	x_1y_4	x_1y_5	x_1y_6	x_1y_7	x_1y_8	x_1y_9
x_2	x_2y_1	x_2y_2	x_2y_3	x_2y_4	x_2y_5	x_2y_6	x_2y_7	x_2y_8	x_2y_9
x_3	x_3y_1	x_3y_2	x_3y_3	x_3y_4	x_3y_5	x_3y_6	x_3y_7	x_3y_8	x_3y_9
x_4	x_4y_1	x_4y_2	x_4y_3	x_4y_4	x_4y_5	x_4y_6	x_4y_7	x_4y_8	x_4y_9
x_5	x_5y_1	x_5y_2	x_5y_3	x_5y_4	x_5y_5	x_5y_6	x_5y_7	x_5y_8	x_5y_9
x_6	x_6y_1	x_6y_2	x_6y_3	x_6y_4	x_6y_5	x_6y_6	x_6y_7	x_6y_8	x_6y_9
x_7	x_7y_1	x_7y_2	x_7y_3	x_7y_4	x_7y_5	x_7y_6	x_7y_7	x_7y_8	x_7y_9
x_8	x_8y_1	x_8y_2	x_8y_3	x_8y_4	x_8y_5	x_8y_6	x_8y_7	x_8y_8	x_8y_9
x_9	x_9y_1	x_9y_2	x_9y_3	x_9y_4	x_9y_5	x_9y_6	x_9y_7	x_9y_8	x_9y_9

b	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9
x_1	x_1y_1	x_1y_2	x_1y_3	x_1y_4	x_1y_5	x_1y_6	x_1y_7	x_1y_8	x_1y_9
x_2	x_2y_1	x_2y_2	x_2y_3	x_2y_4	x_2y_5	x_2y_6	x_2y_7	x_2y_8	x_2y_9
x_3	x_3y_1	x_3y_2	x_3y_3	x_3y_4	x_3y_5	x_3y_6	x_3y_7	x_3y_8	x_3y_9
x_4	x_4y_1	x_4y_2	x_4y_3	x_4y_4	x_4y_5	x_4y_6	x_4y_7	x_4y_8	x_4y_9
x_5	x_5y_1	x_5y_2	x_5y_3	x_5y_4	x_5y_5	x_5y_6	x_5y_7	x_5y_8	x_5y_9
x_6	x_6y_1	x_6y_2	x_6y_3	x_6y_4	x_6y_5	x_6y_6	x_6y_7	x_6y_8	x_6y_9
x_7	x_7y_1	x_7y_2	x_7y_3	x_7y_4	x_7y_5	x_7y_6	x_7y_7	x_7y_8	x_7y_9
x_8	x_8y_1	x_8y_2	x_8y_3	x_8y_4	x_8y_5	x_8y_6	x_8y_7	x_8y_8	x_8y_9
x_9	x_9y_1	x_9y_2	x_9y_3	x_9y_4	x_9y_5	x_9y_6	x_9y_7	x_9y_8	x_9y_9

c	y_2	y_4	y_5	y_1	y_3	y_6	y_7	y_8	y_9
x_3	x_3y_2	x_3y_4	x_3y_5	x_3y_1	x_3y_3	x_3y_6	x_3y_7	x_3y_8	x_3y_9
x_6	x_6y_2	x_6y_4	x_6y_5	x_6y_1	x_6y_3	x_6y_6	x_6y_7	x_6y_8	x_6y_9
x_8	x_8y_2	x_8y_4	x_8y_5	x_8y_1	x_8y_3	x_8y_6	x_8y_7	x_8y_8	x_8y_9
x_1	x_1y_2	x_1y_4	x_1y_5	x_1y_1	x_1y_3	x_1y_6	x_1y_7	x_1y_8	x_1y_9
x_2	x_2y_2	x_2y_4	x_2y_5	x_2y_1	x_2y_3	x_2y_6	x_2y_7	x_2y_8	x_2y_9
x_4	x_4y_2	x_4y_4	x_4y_5	x_4y_1	x_4y_3	x_4y_6	x_4y_7	x_4y_8	x_4y_9
x_5	x_5y_2	x_5y_4	x_5y_5	x_5y_1	x_5y_3	x_5y_6	x_5y_7	x_5y_8	x_5y_9
x_7	x_7y_2	x_7y_4	x_7y_5	x_7y_1	x_7y_3	x_7y_6	x_7y_7	x_7y_8	x_7y_9
x_9	x_9y_2	x_9y_4	x_9y_5	x_9y_1	x_9y_3	x_9y_6	x_9y_7	x_9y_8	x_9y_9

Figure 7. (a) A dimer library can be represented by a 2×2 matrix where the rows of the matrix represent the reactants in one pool and the columns represent the reactants in the other pool. The elements of the matrix represent dimers. The shaded elements represent an example of a subset library, L_d , of the nine most diverse compounds chosen by applying DBCS to the enumerated library, C . The synthetic inefficiency of the method is highlighted by the number of reactants that are required to build the compounds, i.e., x_3 , x_4 , x_5 , x_6 , x_7 , and x_8 are required from pool R_1 and reactants y_1 , y_2 , y_4 , y_6 , y_7 , y_8 , and y_9 are required from pool R_2 . (b) A n_1n_2 subset of the library that is also a combinatorial library can be selected by intersecting n_1 rows with n_2 columns, for example, the 3×3 library built from reactants x_3 , x_6 , and x_8 reacted with reactants y_2 , y_4 , and y_5 is represented by the shaded elements of the matrix. (c) Reordering of the rows and columns of the matrix results in the combinatorial library occupying the top left hand corner of the matrix. Selecting a maximally diverse combinatorial library is then equivalent to reordering the rows and columns of the matrix in order to maximize the diversity of the molecules in the top left-hand corner of the matrix.

not represent combinatorial libraries, and thus it is of limited use in practical combinatorial chemistry.

The synthetic inefficiency resulting from performing DBCS at the product level is illustrated in Figure 7a. A fully enumerated combinatorial library, C , built from two

reactant pools can be visualized as a two-dimensional matrix. The rows of the matrix represent the N_1 reactants available in pool R_1 , and the columns of the matrix represent the N_2 reactants in pool R_2 . The elements of the matrix then represent the full combinatorial library (C), of size N_1N_2 , that would result from reacting all the reactants in R_1 with all the reactants in R_2 . In Figure 7a, pool R_1 contains the nine reactants labeled $x_1 \dots x_9$ and pool R_2 contains the nine reactants $y_1 \dots y_9$. Assume that we wish to select the nine most diverse compounds from C . The DBCS algorithm can select compounds from anywhere within the matrix, for example, the library, L_d^* , selected using DBCS might correspond to the shaded elements, as shown. The nine compounds illustrated require six reactants from pool R_1 and seven reactants from pool R_2 , rather than three from each pool as would be required to build a nine-member subset that is a combinatorial library.

In the experiments performed using the amide library, the 1600 molecules selected from the full amide library are constructed from 137 amines and 146 carboxylic acids. The systematic joining of all these amines to all of the carboxylic acids as performed in practical combinatorial synthesis would result in 19 992 molecules, of which the 1600 most diverse molecules are a subset. The synthetic inefficiency of performing selection at the product level has also been noted by Cribbs *et al.*,²⁶ in their work, nearly all of the reactants were required in order to build the molecules selected. In this section we investigate whether it is possible to generate a combinatorial library from the products that is more diverse than the library generated by selecting at the reactant level.

A nine-member subset of the dimer library, C , that represents a combinatorial library can be selected by intersecting three rows of the matrix with three columns. For example, a 3×3 library built from reactants x_3, x_6 , and x_8 reacted with reactants y_2, y_4 , and y_5 is shown by the shaded elements of the matrix in Figure 7b. For ease of visualization, assume that the rows and columns of the matrix are reordered so that the shaded elements occupy the top left-hand corner of the matrix, Figure 7c, and that the diversity of this sublibrary is then measured. It is possible to reorder the rows and columns so that all possible combinatorial sublibraries can be positioned in the top left-hand corner. Thus, selecting a combinatorial library from product space can be visualized as the reordering of an n -dimensional matrix, where there are n reactant pools involved in the reaction. Finding a maximally diverse combinatorial library is then equivalent to reordering the rows and columns of the matrix and measuring the diversity of all possible sublibraries that occupy the n_1n_2 -member, top left-hand corner of the matrix. Exploring all permutations of rows and columns represents an enormous search space even for libraries of moderate size, and hence, in practice, the manipulation of the matrix is achieved using a genetic algorithm. Simulated annealing has also been applied to the problem of row-column matrix manipulation in a different context.²⁷

The GA is similar to that described in the previous section that was designed to select a maximally diverse subset of reactants, although in this case each chromosome of the GA represents one combinatorial library. For a dimer library of size n_1n_2 a chromosome consists of two parts; the first part represents the n_1 reactants selected from pool R_1 (or the rows of the matrix) and the second part consists of the n_2 reactants

selected from pool R_2 (or the columns of the matrix). The fitness function of the GA is applied to each chromosome and involves constructing the n_1n_2 combinatorial library, C^* , represented by the chromosome and measuring its diversity, $D(C^*)$. Diversity is measured by summing the pairwise dissimilarities as described previously and the GA attempts to maximize the diversity, $D(C^*)$. The GA uses a population of 100 chromosomes. One of the chromosomes in the initial population is initialized to the reactant subsets found by performing DBCS on the reactant pools themselves and thus represents a solution with diversity $D(c)$. The remaining chromosomes are initialized to random subsets.

The genetic operators of one-point crossover and mutation are implemented. In each case, duplicate entries in either half of the chromosome are forbidden. Mutation involves altering some elements of the chromosome to new elements. The mutation operation is equivalent to exchanging a reactant in one selected subset with a different one from the relevant pool. This is equivalent to exchanging a row or a column from the top left-hand corner of the matrix with a row or a column that appears lower in the matrix. Crossover creates two new child chromosomes and is applied to one part of the parent chromosome only, *i.e.*, to one of the reactant pools. A crossover point is chosen at random, and the reactants in that subset in each parent are exchanged after that crossover point, provided that they do not result in duplicate entries in either half of the chromosome. This is equivalent to having one of the subsets remain unchanged in each parent and mixing the reactants between the parents in the other subsets. As in the previous GA, the mutation operator is applied at a higher rate than crossover (3:1) in order to reduce the chances of the GA finding a local rather than a global maximum. The GA is a steady-state with no-duplicates algorithm.

The GA was run on the libraries described previously. The best results were obtained by repeatedly rerunning the GA with a new starting population, consisting of the best 10% of the chromosomes from the previous run, with the remaining chromosomes initialized to random subsets. The reactant pools each consisted of 400 reactants which when reacted together generated libraries (C) of size 160 000. DBCS was performed at the product level to select libraries L_d^* of size 1600. DBCS was also performed at the reactant level to generate libraries, c , also containing 1600 molecules. Finally, combinatorial libraries, C^* , were selected from the products and their diversities measured and compared with the $D(c)$ and $D(L_d^*)$.

The results obtained are shown in Table 7. In all cases combinatorial libraries are selected that are significantly more diverse than if compound selection is performed at the reactant level. In fact, the combinatorial libraries are intermediate in diversity between performing DBCS on the reactants and performing DBCS on the products. Although the combinatorial libraries selected using the GA are not as diverse as the DBCS libraries that are selected from product space, they are synthetically efficient, and the reagents represented by them can be fed directly into combinatorial synthesis experiments. This algorithm is therefore a significant improvement over performing DBCS on the reactant pools. Not only is it highly effective in operation but it is also very efficient, with the selection of 40×40 reactant pools from a 400×400 virtual library requiring ap-

Table 7. Comparison of Results from Applying DBCS to the Reactants and Enumerating Libraries $D(c)$, Applying DBCS to the Fully Enumerated Product Library $D(L_d)$ and Using the GA To Select Diverse Combinatorial Libraries from the Products, $D(C^*)$

library	#(c)	$D(L_d^*)$	$D(C^*)$	$D(c)$
amides	1600	0.652	0.623	0.596
	900	0.658	0.628	0.594
	400	0.663	0.632	0.596
	100	0.674	0.637	0.582
benzoic acid	1600	0.408	0.396	0.388
	900	0.410	0.396	0.387
	400	0.413	0.398	0.384
	100	0.421	0.399	0.373
Kemp's acid	1600	0.435	0.419	0.401
	900	0.436	0.422	0.401
	400	0.440	0.422	0.403
	100	0.448	0.419	0.396

proximately 20 min using a Silicon Graphics R10000 processor.

DISCUSSION

The diversity hypothesis has been tested for three different libraries and using two different structural representations. In all cases it is seen that more diverse libraries result from applying DBCS at the product level rather than at the reactant level. Also the experiments designed to find the maximally diverse subset suggest that DBCS is near-optimal. It is concluded that the diversity hypothesis is not supported and that there is still considerable scope for improving the diversity of libraries resulting from applying compound selection to reactant pools.

A significant limitation of performing DBCS at the product library level is that the resulting libraries do not represent "true" combinatorial libraries, that is, they require larger pools of reactants than if selection is performed at the reactant level, and the reactants are enumerated into products in all possible ways. We have hence developed an algorithm for selecting combinatorial libraries from product libraries that are significantly more diverse than if selection is performed by analyzing reactant space. This is a practical solution that allows reactants to be selected by analyzing product space. The GA has been applied to optimize subset diversity according to one measure, that is, the sum of pairwise dissimilarities of the selected molecules. However, it could also be applied to other diversity measures, for example, partition-based measures that look to maximize the coverage of partitions by the selected molecules, if these could be performed sufficiently rapidly to form the fitness function for the GA.

The limitations of the experiments must be emphasized. Firstly, the libraries considered contain just 160 000 molecules, whereas much larger virtual libraries are easily conceived. Although the libraries studied represent different types of "reactions," ranging from a library where all of the products share a large common core substructure, they were all based (in whole or in part) on carboxylic acid reactants, and it may be that different types of reactants lead to different results.

Secondly, only one kind of subset-selection method was considered, that of dissimilarity-based compound selection. Other methods of selecting subsets include clustering, partition-based selection, and D-optimal design. However,

these methods are much more computationally intensive, and it is not practical to apply them to large collections of compounds. For example, Cribbs *et al.*²⁶ compared selection at the virtual library level with reactant-based selection using D-optimal design, clustering, and a uniform shell approach. They concluded that the computational costs of selecting even 900 compounds from 14 000 using these methods were too high.

Another limitation of these experiments is that only one type of diversity measure has been considered; several other such measures have been reported,^{6,9,16,28} and some of these alternatives might be applied to the testing of the diversity hypothesis or might be used as the fitness function of the GA.

In conclusion, this paper has discussed the use of dissimilarity-based compound selection (DBCS) to provide a quantitative test of the diversity hypothesis. Experiments with several different combinatorial libraries show that reactant-based selection procedures result in sets of products that are intermediate in diversity between comparably-sized sets selected at random from a fully enumerated product library and selected from that library to maximize diversity. Thus, while reactant-based selection is an extremely efficient way of generating combinatorial libraries, it is less effective than, ideally, one might wish. Our results also suggest that existing algorithms for DBCS identify subsets that are very close in diversity to optimally dissimilar subsets. Finally, we have described an algorithm for selecting combinatorial libraries from enumerated product libraries that are significantly more diverse than those generated using reactant-based selection.

ACKNOWLEDGMENT

We thank the following: GlaxoWellcome Research and Development for funding; Daylight Chemical Information Systems Inc. for software support; Stan Young for bringing extreme value methods to our attention and Clive Anderson for help with their implementation; Sharon Dankwardt for prepublication details of her work on 3-amino-5-hydroxybenzoic acid libraries; and Darren Green for helpful discussions on selecting combinatorial libraries from product space. The Krebs Institute for Biomolecular Research is a designated Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

REFERENCES AND NOTES

- (1) Gallop, M. A.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. Applications of Combinatorial Technologies to Drug Discovery. 1. Background and Peptide Combinatorial Libraries. *J. Med. Chem.* 1994, 37, 1233-1251.
- (2) Gordon, E. M.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gallop, M. A. Applications of Combinatorial Technologies to Drug Discovery. 2. Combinatorial Organic Synthesis, Library Screening Strategies, and Future Directions. *J. Med. Chem.* 1994, 37, 1385-1399.
- (3) Zuckermann, R. N.; Martin, E. J.; Spellmeyer, D. C.; Stauber, G. B.; Shoemaker, K. R.; Kerr, J. M.; Figliozzi, G. M.; Goff, D. A.; Siani, M. A.; Simon, R. J.; Banville, S. C.; Brown, E. G.; Wang, L.; Richter, L. S.; Moos, W. H. Discovery of Nanomolar Ligands for 7-trans-membrane G-protein-coupled Receptors from a Diverse n-(substituted)-glycine Peptoid Library. *J. Med. Chem.* 1994, 37, 2678-2685.
- (4) *Combinatorial Chemistry*; Czarnik, A. W., DeWitt, S. H., Eds.; American Chemical Society: Washington, DC, 1997.
- (5) Sheridan, R. P.; Kearsley, S. K. Using a Genetic Algorithm to Suggest Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* 1995, 35, 310-320.
- (6) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity; Experimental Design of

- Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* 1995, 38, 1431-1436.
- (7) Downs, G. M.; Willett, P. In *Advanced Computer-Assisted Techniques in Drug Discovery*; van de Waterbeemd, H., Ed.; VCH: New York, 1994; pp 111-130.
- (8) Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Comput. Sci.* 1995, 35, 59-67.
- (9) Shemetulskis, N. E.; Dunbar, J. B.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. Enhancing the Diversity of a Corporate Database Using Chemical Database Clustering and Analysis. *J. Comput.-Aided Mol. Design* 1995, 9, 407-416.
- (10) Hudson, B. D.; Hyde, R. M.; Raha, E.; Wood, J. Parameter Based Methods for Compound Selection from Chemical Databases. *Quant. Struct.-Act. Relat.* 1996, 15, 285-289.
- (11) Lajiness, M. In *QSAR: Rational Approaches to the Design of Bioactive Compounds*; Silipo, C., Vittoria, A., Eds.; Elsevier Science Publishers: Amsterdam, 1991; pp 201-204.
- (12) Ferguson, A. M.; Patterson, D. E.; Garr, C. D.; Underiner, T. L. Designing Chemical Libraries for Lead Discovery. *J. Biomolec. Screen.* 1996, 1, 65-73.
- (13) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* 1996, 36, 572-584.
- (14) Holliday, J. D.; Ranade, S. S.; Willett, P. A Fast Algorithm for Selecting Sets of Dissimilar Molecules from Large Chemical Databases. *Quant. Struct.-Act. Relat.* 1995, 14, 501-506.
- (15) James, C. A.; Weininger, D. *Daylight Theory Manual*; Daylight Chemical Information Systems Inc.: Irvine, CA, 1995.
- (16) Turner, D. B.; Tyrrell, S. M.; Willett, P. Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *J. Chem. Inf. Comput. Sci.* 1997, 37, 18-22.
- (17) WDI is maintained by Derwent Information Ltd., London.
- (18) Dankwardt, S. M.; Phan, T. M.; Krstenansky, J. L. Combinatorial Synthesis of Small-Molecule Libraries Using 3-amino-5-hydroxybenzoic Acid. *Molec. Diversity* 1995, 1, 113-120.
- (19) The SPRESI database is distributed by Daylight Chemical Information Systems, Inc., Irvine, CA.
- (20) Kocis, P.; Issakova, O.; Sepetov, N. F.; Lebl, M. Kemp's Triacid Scaffolding for Synthesis of Combinatorial Nonpeptide Uncoded Libraries. *Tetrahedron Lett.* 1995, 36, 6623-6626.
- (21) Reeves, C. R. *Modern Heuristic Techniques for Combinatorial Problems*; McGraw-Hill Book Company: Europe, 1995.
- (22) Goldberg, D. E. *Genetic Algorithms in Search, Optimisation, and Machine Learning*; Addison-Wesley: Reading, MA, 1989.
- (23) Clark, D. E.; Westhead, D. R. Evolutionary Algorithms in Computer-Aided Molecular Design. *J. Comput.-Aid. Mol. Design* 1996, 10, 337-358.
- (24) Coles, S. G.; Tawn, J. A. Statistical-Methods for Multivariate Extremes - An Application to Structural Design. *Appl. Statist.* 1994, 43, 1-48.
- (25) Davidson, A. C.; Smith, R. L. Models for Exceedances over High Thresholds. *J. Roy. Stat. Soc. B52*, No. 3, 1990, 393-442.
- (26) Cribbs, C. M.; Menius, J. A.; Cummins, D.; Young, S. S. Statistical-Methods for Monomer Selection in Chemical Library Design. *Abstracts of Papers of the American Chemical Society*; 1996; Vol. 211, No. Part 1, pp 74-COMP.
- (27) Packer, C. V. Applying Row-Column Permutation to Matrix Representations of Large Citation Networks. *Information Processing and Management* 1989, 25, 307-314.
- (28) Boyd, S. M.; Beverley, M.; Norskov, L.; Hubbard, R. E. Characterizing the Geometric Diversity of Functional-Groups in Chemical Databases. *J. Comput.-Aid. Mol. Design* 1995, 9, 417-424.

CI970420G